Escaping the Token Trap: Joint Embedding Predictive Architectures and the Paradigm Shift Toward AI World Models

Douglas C. Youvan

doug@youvan.com

www.youvan.ai

March 11, 2026

The rapid advancement of artificial intelligence has been largely driven by autoregressive large language models operating via next-token prediction in the input space. While these generative models excel at syntactic fluency and pattern recognition, their reliance on reconstructing discrete tokens or raw pixels leaves them fundamentally limited in grasping physical reality, causality, and complex temporal reasoning. This paper explores the paradigm shift championed by the Joint Embedding Predictive Architecture and the pursuit of artificial intelligence world models. By shifting the predictive objective from a high-dimensional input space to an abstract and continuous latent space, this alternative architecture bypasses the computational inefficiencies of predicting irrelevant microscopic details. Instead, it filters unpredictable noise to capture high-level semantic structures, enabling systems to simulate physical dynamics, handle uncertainty gracefully, and perform complex planning. As artificial intelligence expands into robotics, autonomous systems, and industrial automation, the transition from generative token prediction to discriminative latent space prediction represents a critical evolution. This comprehensive analysis contrasts the architectural foundations of standard generative models with latent predictive models, examines the underlying mechanics of abstract state prediction, and evaluates the implications of world models for the future of machine intelligence.

Joint Embedding Predictive Architecture, JEPA, World Models, Artificial Intelligence, Large Language Models, Latent Space Prediction, Next-Token Prediction, Autonomous Agents, Causality, System 2 AI.

A collaboration with Google Gemini-3.1-Pro. CC4.0.

Outline

1. Introduction 1.1. The Current Landscape of Artificial Intelligence 1.2. The Emergence of Advanced Machine Intelligence and World Models 1.3. Scope and Structure of the Paper

2. The Limitations of Next-Token Prediction 2.1. Autoregressive Architectures and the Input Space 2.2. The Computational Burden of Pixel and Token Reconstruction 2.3. The Absence of Physical Grounding and Causality

3. Defining the Paradigm of AI World Models 3.1. Conceptual Origins and Theoretical Framework 3.2. Cognitive Simulation versus Statistical Memorization 3.3. Goal-Oriented Planning and Sequence Generation

4. The Architectural Foundation of JEPA 4.1. Core Components of the Joint Embedding Predictive Architecture 4.2. The Role of the Encoder and Predictor Networks 4.3. Contrastive versus Non-Contrastive Learning Approaches

5. Input Space versus Latent Space 5.1. The Mathematical Shift to Continuous Embeddings 5.2. Dimensionality Reduction and Semantic Extraction 5.3. Filtering Irrelevant Micro-Details from Core Events

6. Handling Uncertainty in Machine Learning 6.1. The Hallucination Problem in Generative Models 6.2. Embracing Aleatoric Uncertainty through Latent Abstraction 6.3. Predicting Abstract States in Unpredictable Environments

7. Generative versus Predictive Discriminative Models 7.1. The Objective Function Dilemma 7.2. Decoding Latent Vectors for Human-Readable Outputs 7.3. Energy-Based Models as a Unifying Theory

8. Training Methodologies and Variants 8.1. Image-Based Learning with I-JEPA 8.2. Temporal Dynamics and Masked Video Prediction with V-JEPA 8.3. Audio and Multimodal Expansions

9. Applications in Robotics and Autonomous Systems 9.1. Navigating Continuous Physical Environments 9.2. Industrial Automation and Real-Time Decision Making 9.3. Simulation Engines for Autonomous Driving

10. Integrating System 2 Thinking in Artificial Intelligence 10.1. Defining Fast versus Slow Thinking in Neural Networks 10.2. Logical Reasoning and Counterfactual Simulation 10.3. Escaping the Reactive Feedback Loop

11. Challenges and Scalability 11.1. Hardware and Training Complexities 11.2. Representation Collapse and Mitigation Strategies 11.3. The Future Trajectory of Joint Embedding Architectures

12. Conclusion 12.1. Summary of Architectural Contrasts 12.2. Final Thoughts on the Future of World Models 12.3. Call for Future Research

References

Art

1. Introduction

1.1. The Current Landscape of Artificial Intelligence Artificial intelligence has undergone a profound transformation over the last decade, primarily driven by the rapid scaling of neural networks and the ingestion of massive internet-scale datasets. The contemporary landscape is heavily dominated by large language models and generative architectures that rely on autoregressive techniques. These systems operate fundamentally by calculating probability distributions to predict the next discrete unit of data, whether that is the next word in a sentence or the next pixel in an image. This methodology, operating entirely within the input space, has yielded unprecedented success in natural language processing and media generation. It has enabled machines to generate highly fluent text, write complex computer code, and produce photorealistic media on demand. However, the field is beginning to recognize that reliance on statistical pattern matching within high-dimensional input spaces carries significant structural limitations. While generative models excel at syntax, mimicry, and surface-level correlation, they lack an intuitive grasp of physical reality, common sense reasoning, and causality. As researchers attempt to deploy artificial intelligence in unpredictable, real-world physical environments, the ceiling of purely autoregressive paradigms has become increasingly apparent. This realization has prompted an industry-wide search for alternative architectures capable of deeper, grounded comprehension.

1.2. The Emergence of Advanced Machine Intelligence and World Models In response to the computational and cognitive bottlenecks of autoregressive generation, a competing paradigm has steadily gained momentum, heavily championed by prominent researchers such as Yann LeCun and foundational to new ventures like Advanced Machine Intelligence. This alternative approach fundamentally rejects the premise that artificial general intelligence can be achieved merely by scaling next-token prediction. Instead, it advocates for the development of world models. A world model is an internal mathematical engine that allows an artificial agent to understand how its environment operates at a structural level. It enables a system to observe a current state, simulate potential future states, and plan a sequence of actions to achieve specific goals, all without needing to calculate or render the microscopic, irrelevant details of the environment. At the heart of this theoretical movement is the Joint Embedding Predictive Architecture. By shifting the machine learning objective from reconstructing raw input data to predicting abstract representations in a continuous latent space, this architecture allows systems to filter out noise and learn the underlying causal structure of reality. This represents a crucial pivot from generative artificial intelligence, which focuses on producing human-consumable media, to objective-driven artificial intelligence designed for physical intuition and temporal reasoning.

1.3. Scope and Structure of the Paper This paper provides a comprehensive examination of the transition from autoregressive input space prediction to latent space prediction via the Joint Embedding Predictive Architecture. The subsequent sections systematically deconstruct the mechanics, advantages, and applications of artificial intelligence world models. Section 2 explores the specific computational and cognitive limitations of current generative architectures. Section 3 explicitly defines the conceptual framework of world models and their role in goal-oriented planning. Sections 4 and 5 delve into the structural foundation of latent space prediction, mathematically and conceptually contrasting it with traditional input space operations. Section 6 addresses how this new paradigm handles uncertainty gracefully without succumbing to the hallucination problem prevalent in generative systems. Section 7 clarifies the vital distinction between generative media models and predictive discriminative models. Sections 8 and 9 examine specific training methodologies, such as video-based prediction, alongside their direct applications in robotics and autonomous systems. Finally, Sections 10 through 12 explore the integration of logical reasoning, current hardware and scalability challenges, and the broader trajectory of machine intelligence. By mapping this architectural shift, this paper elucidates the path toward more robust, physically grounded, and capable artificial systems.

2. The Limitations of Next-Token Prediction

2.1. Autoregressive Architectures and the Input Space Autoregressive models operate by predicting future values based on a sequence of past observations. In the context of large language models and generative artificial intelligence, this translates to estimating the conditional probability of the next discrete unit of data given all preceding units, expressed mathematically as $P(x_t \mid x_1, x_2, ..., x_{t-1})$. Crucially, this operation occurs entirely within the input space. The input space consists of the raw, uncompressed data formats directly consumed and produced by the system, such as text tokens, individual image pixels, or raw audio waveforms.

By forcing the model to calculate probabilities and generate outputs at this highly granular level, the architecture inherently prioritizes surface-level pattern replication over deep structural comprehension. The model learns the statistical likelihood of a specific token or pixel following a given sequence, which yields remarkable syntactic fluency and visual mimicry but fundamentally fails to capture the underlying semantic, logical, or physical principles governing the sequence.

2.2. The Computational Burden of Pixel and Token Reconstruction Operating directly in the input space imposes an extraordinary computational burden, particularly as machine learning

models expand beyond text into continuous, multimodal applications like spatial simulation and video generation. When a generative model attempts to predict the next frame of a high-resolution video, it is tasked with reconstructing millions of individual pixels. This requires the neural network to expend immense computational resources predicting highly stochastic, irrelevant micro-details that have no bearing on the actual core event taking place. For instance, if a video depicts a vehicle driving down a road, an input space model must dedicate capacity to rendering the exact random movement of leaves on a distant tree in the wind, the precise microscopic texture of the asphalt, and the inherent noise in the recording camera sensor. Because the objective function penalizes the model for any deviation from the raw pixel data, the system is forced to model unpredictable noise alongside the essential signal. This inefficiency creates a severe scaling bottleneck. As the resolution, complexity, and temporal duration of the data increase, the computational power required to predict every atomic unit scales exponentially, making it practically infeasible to simulate long-term, continuous futures in real-world environments.

2.3. The Absence of Physical Grounding and Causality Perhaps the most significant limitation of next-token prediction is its fundamental inability to develop a grounded understanding of physical reality and true causality. Because autoregressive models are trained purely to minimize loss based on statistical correlations within vast static datasets, they do not learn how the physical world actually operates; they merely learn how humans historically write about or visually record the world. An autoregressive language model can articulate that a glass will shatter if dropped onto concrete because those specific linguistic tokens frequently co-occur in its training corpus. However, it possesses no internal, mathematical representation of gravity, object permanence, mass, velocity, or structural fragility. It cannot simulate the physical dynamics of the event. Consequently, when these generative models are confronted with novel, out-of-distribution physical scenarios that deviate slightly from their training data, they frequently fail, producing physically impossible, hallucinated, or logically inconsistent outputs. This strict absence of causal reasoning renders traditional generative models fundamentally inadequate for integration into autonomous agents, robotics, and automated systems that must plan, interact, and operate safely within the continuous physical world.

3. Defining the Paradigm of AI World Models

3.1. Conceptual Origins and Theoretical Framework The concept of a world model draws deep inspiration from cognitive science, neuroscience, and control theory. In human cognition, the brain does not merely react to sensory stimuli; it continuously constructs and updates an internal mental model of the environment. This mental model allows humans to anticipate the consequences of their actions, infer missing information, and navigate complex physical spaces

without needing to consciously process every photon hitting the retina. In the context of artificial intelligence, a world model is a mathematical engine designed to replicate this predictive capability. The theoretical framework posits that an intelligent agent must possess an internal representation of its environment to function autonomously. Formally, a world model learns a transition function where the future state of the environment is predicted based on the current state and a proposed action, often represented mathematically as $s_{t+1} = f(s_t, a_t)$. By decoupling the learning process from raw sensory inputs and focusing on the underlying dynamics of the environment, a world model provides a structured, causal understanding of how the world evolves over time.

3.2. Cognitive Simulation versus Statistical Memorization To understand the paradigm shift, it is essential to distinguish between cognitive simulation and statistical memorization. Traditional autoregressive models function primarily as highly sophisticated retrieval and memorization systems. They map inputs to outputs based on the statistical distribution of vast training datasets. If a specific pattern was well-represented in the training data, the model can reproduce it; if the pattern requires novel physical reasoning, the model falters. Conversely, a world model acts as a cognitive simulator. Rather than memorizing the precise outcome of a specific event, it learns the underlying rules governing the event. For example, instead of memorizing the pixel-by-pixel trajectory of a bouncing ball, a world model learns the abstract principles of momentum, gravity, and elasticity. When presented with a novel scenario, the world model does not search its memory for a statistical match. Instead, it initializes a simulation based on the current abstract state and runs the learned dynamic rules forward in time. This enables the artificial intelligence to generalize to entirely unseen situations, providing a foundation for intuitive physics and common sense reasoning that statistical memorization simply cannot achieve.

3.3. Goal-Oriented Planning and Sequence Generation The ultimate utility of a world model lies in its capacity for goal-oriented planning and sequence generation. In reinforcement learning and autonomous control, an agent must often execute a long sequence of actions to achieve a specific objective, such as a robotic arm navigating an obstacle course to grasp an object. Without a world model, an agent must rely on trial and error in the physical world, which is slow, dangerous, and computationally expensive. With a world model, the agent can perform these trials internally. This process is analogous to human deliberation or what cognitive psychologists refer to as System 2 thinking. The agent can imagine multiple different action sequences, simulate the trajectory of the environment for each sequence using its internal world model, and evaluate the predicted outcomes against a predefined cost function or goal state. By searching through these simulated futures, the agent can identify the optimal sequence of actions before making a single move in physical reality. This internal planning

mechanism is the cornerstone of advanced machine intelligence, separating purely reactive systems from truly deliberate, intelligent agents.

4. The Architectural Foundation of JEPA

4.1. Core Components of the Joint Embedding Predictive Architecture The Joint Embedding Predictive Architecture represents a structural departure from traditional generative models by prioritizing the prediction of abstract representations over the reconstruction of raw sensory inputs. At its core, the architecture consists of distinct neural network modules designed to process and predict information entirely within a continuous, mathematically defined latent space. When presented with a data sample, the system partitions the input into a visible context segment and an obscured target segment. The overarching objective is to utilize the available context to predict the underlying semantic structure of the unseen target. Unlike masked autoencoders that attempt to fill in missing pixels or text tokens directly in the input space, this architecture maps both the context and the target into an embedded representation before any prediction occurs. This deliberate decoupling of the predictive mechanism from the high-dimensional input space is the defining characteristic of the architecture. It enables the model to ignore stochastic, unpredictable noise and focus its computational resources entirely on the deterministic, high-level features of the environment.

4.2. The Role of the Encoder and Predictor Networks The functionality of this architecture relies heavily on the synergistic relationship between its encoder and predictor networks. The encoder is a neural network responsible for transforming high-dimensional raw data into a compressed, abstract latent vector. Mathematically, if we define a context input as x and a target input as y, the encoder function E maps these inputs to their respective latent states, such that s_x = E(x) and s_y = E(y). Once the context is encoded into the state s_x, it is fed into the predictor network. The predictor function P takes this context representation, often alongside a condition or latent variable z that accounts for action or uncertainty, to generate a prediction of the target's latent state. This dynamic relationship is expressed as s_hat_y = P(s_x, z).

The system is trained by minimizing the distance or divergence between the predicted state s_hat_y and the actual encoded target state s_y. By restricting the prediction task entirely to the abstract outputs of the encoder, the predictor network is freed from the computational burden of rendering exact spatial or temporal details. This allows the network to dedicate its capacity to learning the fundamental dynamic transitions and semantic relationships between different states.

4.3. Contrastive versus Non-Contrastive Learning Approaches A significant historical challenge in training joint embedding architectures is the phenomenon of representation collapse. This occurs when the encoder trivially learns to map all distinct inputs to a single constant vector, thereby achieving zero prediction error without extracting any meaningful features from the data. Previously, contrastive learning methods mitigated this issue by artificially pushing the embeddings of different samples apart in the latent space while pulling augmented versions of the same sample together. However, contrastive approaches are notoriously computationally expensive and struggle to scale efficiently across highly complex or continuous multimodal environments. The Joint Embedding Predictive Architecture bypasses this limitation by utilizing non-contrastive learning techniques. Instead of relying on vast batches of negative samples, the architecture employs an asymmetric design to prevent collapse. Typically, the target encoder operates as an exponential moving average of the context encoder, meaning its parameters are updated smoothly over time rather than via direct gradient descent. This stop-gradient mechanism ensures that the target representations remain stable and information-rich, compelling the predictor network to genuinely learn the underlying structure of the data rather than exploiting trivial mathematical solutions. This non-contrastive foundation is critical for scaling world models to handle the vast state spaces encountered in real-world physical systems.

5. Input Space versus Latent Space

5.1. The Mathematical Shift to Continuous Embeddings The fundamental distinction between autoregressive generative models and the Joint Embedding Predictive Architecture lies in the mathematical space where prediction occurs. Input space refers to the raw, observable data format, such as the discrete vocabulary tokens of a language model or the multi-channel pixel grids of an image. In mathematical terms, an image in the input space is typically represented as a high-dimensional tensor, such as x in R^(H x W x C), representing height, width, and color channels. Operating in this space forces a model to predict the exact sensory observation. In contrast, latent space is a hidden, continuous mathematical domain where data is represented as dense vectors, or embeddings, typically denoted as z in R^d, where d is the number of dimensions. The shift from predicting in the input space to predicting in the latent space means the model is no longer forecasting the exact sensory experience but rather the underlying mathematical representation of that experience. Because the latent space is continuous, it allows for smooth interpolations and geometric relationships that discrete input spaces cannot support, enabling the system to map conceptual similarities and physical laws directly into the geometry of the embedding space.

5.2. Dimensionality Reduction and Semantic Extraction Transitioning from the input space to the latent space fundamentally relies on dimensionality reduction, a process where a vast amount of raw data is compressed into a significantly smaller, more manageable format. In this architecture, the latent dimension is vastly smaller than the input dimension. This compression is not merely a method of saving computational memory; it is the primary mechanism for semantic extraction. An encoder network acts as an information bottleneck. By forcing high-dimensional data, such as a high-resolution video frame, through this bottleneck, the network is mathematically compelled to discard superficial variations and retain only the most critical, defining features of the data.

The resulting latent vector is a concentrated distillation of meaning. It captures the essence of the objects, their spatial relationships, and their contextual significance without retaining the exact pixel values, textures, or lighting conditions that generated the original image. This extraction of semantics ensures that the predictive architecture operates entirely on the level of concepts and physical states rather than raw sensory inputs.

5.3. Filtering Irrelevant Micro-Details from Core Events The most practical advantage of operating exclusively within a semantic latent space is the inherent ability to filter out unpredictable and irrelevant micro-details from core events. The physical world is filled with stochastic noise that is impossible to predict accurately and ultimately useless for high-level planning. If a system is tasked with predicting the outcome of a vehicle driving down a gravel road, an input space model must attempt to render the exact trajectory of every pebble kicked up by the tires and the precise movement of clouds in the sky. These micro-details are computationally expensive to calculate and largely irrelevant to the core event of navigation. A latent space model, utilizing the Joint Embedding Predictive Architecture, bypasses this problem entirely. The encoder maps the complex visual of the moving car into a stable latent state that simply represents the vehicle's velocity, trajectory, and immediate obstacles. The predictor network only needs to forecast this abstract transition. By stripping away the noise of the physical world, the model can focus its computational resources on tracking the causal, macroscopic events that actually matter for decision making and autonomous action.

6. Handling Uncertainty in Machine Learning

6.1. The Hallucination Problem in Generative Models Generative architectures are fundamentally constrained by their requirement to output a specific point in a high-dimensional input space. When a generative model is uncertain about the future, such as predicting what exists behind a moving object in a video sequence or forecasting the physical trajectory of a complex mechanism, it faces a multimodal distribution of possible realities. If forced to predict

exact pixel values under this uncertainty, the model will typically attempt to minimize its error by averaging all possible outcomes. Visually, this manifests as a blurry, amorphous, and physically impossible rendering because the mathematical average of several distinct physical states is rarely a valid physical state itself. Alternatively, the model might arbitrarily sample one specific path from the distribution, resulting in a confident but entirely fabricated output. In natural language processing, this phenomenon is widely recognized as a hallucination, where a model generates plausible-sounding but factually incorrect or logically inconsistent statements. The root cause of these failures is the structural mismatch between the unpredictable, branching nature of physical reality and the rigid architectural mandate to reconstruct exact input tokens or pixels. Because they lack a mechanism to represent ambiguity gracefully, generative models are forced into premature and often incorrect commitments.

6.2. Embracing Aleatoric Uncertainty through Latent Abstraction Uncertainty in physical environments is frequently aleatoric, meaning it is inherent, irreducible, and driven by the natural stochasticity of the world. The Joint Embedding Predictive Architecture explicitly addresses this challenge by shifting the prediction target from the deterministic input space to a flexible, continuous latent space. To handle branching possibilities, the architecture introduces a latent variable, conventionally denoted as z, which represents the unknown or unpredictable factors of the environment. The predictor network utilizes this variable alongside the encoded context state to forecast the future, modifying the predictive function to s_hat_y = P(s_x, z).

By drawing different samples for the latent variable z from a predefined probability distribution, the predictor network can generate a manifold of plausible future states rather than a single, compromised average. This structural inclusion of a latent variable allows the architecture to acknowledge, quantify, and mathematically represent multiple valid realities simultaneously. The model learns to map the predictable elements of the environment deterministically while using the latent variable to absorb and account for the irreducible noise, entirely avoiding the need to hallucinate specific details.

6.3. Predicting Abstract States in Unpredictable Environments In real-world applications such as autonomous driving, drone navigation, or advanced robotics, the physical environment is inherently chaotic. A pedestrian might subtly alter their gait, a gust of wind might scatter leaves across a sensor, or the precise reflection of sunlight on a wet road might shift unpredictably. A functional artificial intelligence world model must navigate this chaos without allowing its internal simulations to collapse or diverge wildly. Because the Joint Embedding Predictive Architecture predicts abstract conceptual states rather than literal sensory inputs, it naturally filters out the unpredictable, high-frequency noise that disrupts generative models. If an autonomous agent observes a complex intersection, the exact pixel-level trajectory of every distant vehicle and pedestrian is highly uncertain and computationally prohibitive to forecast.

Instead of attempting to hallucinate the micro-details of this scene, the latent world model predicts the overarching abstract semantic continuity of the traffic flow and the safety state of the immediate path. This capacity to selectively ignore irreducible uncertainty while accurately forecasting predictable, macroscopic events is what makes latent space prediction exceptionally robust for long-horizon planning and temporal reasoning in highly dynamic physical settings.

7. Generative versus Predictive Discriminative Models

7.1. The Objective Function Dilemma The architectural divergence between traditional artificial intelligence and world models is most clearly articulated through their respective objective functions. Generative models are bound by a generative objective, which mandates the reconstruction or prediction of raw sensory data. Whether utilizing cross-entropy loss for discrete text tokens or mean squared error for continuous pixel intensities, the optimization process forces the network to model the complete data distribution. Mathematically, a generative loss function such as $L_{gen} = ||x - \hat{x}||^2$ penalizes the model for any deviation in the input space, compelling it to dedicate vast computational resources to capturing stochastic noise and irrelevant background details. Conversely, predictive discriminative architectures like the Joint Embedding Predictive Architecture utilize a latent objective function. The network is trained to minimize the distance between a predicted abstract state and a true encoded state, expressed as $L_{pred} = ||s_y - \hat{s}_y||^2$. By discarding the generative requirement to reconstruct the input x, the discriminative model is entirely liberated from modeling unpredictable noise. It does not generate reality; it discriminates whether a predicted sequence of abstract features logically aligns with the true underlying dynamics of the environment.

7.2. Decoding Latent Vectors for Human-Readable Outputs Because predictive discriminative models operate exclusively within a compressed latent space, their raw outputs are continuous mathematical vectors rather than human-readable text, images, or audio. A system simulating the physics of a collapsing bridge will output a sequence of dense embeddings representing the structural failure, but it will not natively render a video of the event. To translate these abstract cognitive simulations into a format that humans can consume or evaluate visually, the architecture requires the integration of an auxiliary decoder network. This decoder takes the predicted latent state $\hat{s}_y$ and maps it back into the high-dimensional input space. Crucially, in a properly designed world model, this decoder is entirely decoupled from the core reasoning and planning loop. The artificial agent simulates futures and makes decisions entirely in the fast, efficient latent space. The decoder is utilized strictly as a lightweight visualization tool or an interface layer for specific downstream tasks. This separation of cognitive simulation from

sensory generation ensures that the agent's internal planning process is not bottlenecked by the computationally expensive task of rendering pixels.

7.3. Energy-Based Models as a Unifying Theory The mathematical foundation underpinning these predictive discriminative architectures is best understood through the lens of Energy-Based Models, a theoretical framework heavily utilized by Yann LeCun to describe complex dependencies in machine learning. An Energy-Based Model learns an energy function E(x, y) that assigns a scalar energy value to a pair of variables, such as a context x and a future state y. The system is designed to output a low energy value when the variables are highly compatible or logically consistent, and a high energy value when they are incompatible or physically impossible.

Traditional probabilistic generative models attempt to convert these energy values into normalized probabilities, a process that requires calculating the intractable integral over all possible future states in high-dimensional continuous environments. The Joint Embedding Predictive Architecture bypasses this intractable normalization entirely. It simply minimizes the energy of compatible pairs—ensuring the predicted latent state closely matches the true future state—while relying on architectural constraints, such as asymmetric encoder updates and stop-gradients, to naturally push up the energy of incompatible pairs. This framework provides a robust, scalable method for training artificial intelligence to understand physical compatibility and temporal logic without the mathematical limitations of purely probabilistic generative systems.

8. Training Methodologies and Variants

8.1. Image-Based Learning with I-JEPA The Joint Embedding Predictive Architecture was initially validated in the visual domain through the development of the Image Joint Embedding Predictive Architecture. The primary objective of this model is to learn highly semantic image representations without relying on human-annotated labels or computationally intensive pixel reconstruction. In practice, the system processes an image by dividing it into a grid of non-overlapping patches. A significant portion of these patches is mathematically masked or obscured, while the remaining visible patches serve as the context. The context patches are processed through the encoder network to produce a latent representation. Simultaneously, the masked target patches are processed through a target encoder. The predictor network is then tasked with utilizing the context embedding and the positional information of the target patches to predict their corresponding latent representations. Because the system is heavily penalized for errors in the continuous latent space rather than the discrete pixel space, it inherently learns to group related pixels into coherent objects and understand spatial relationships.

This spatial masking methodology forces the architecture to develop a deep, structural understanding of visual semantics, proving that pixel-level generation is unnecessary for high-level image comprehension.

8.2. Temporal Dynamics and Masked Video Prediction with V-JEPA While static image comprehension is a crucial foundational step, true artificial intelligence world models require an understanding of temporal dynamics and physical causality. This necessity led to the evolution of the Video Joint Embedding Predictive Architecture. Moving beyond two-dimensional spatial masking, this variant introduces a third dimension by masking entire spatiotemporal tubes within a video sequence. The system might observe the first few frames of a video, experience a temporal mask where several subsequent frames are entirely obscured, and then be required to predict the latent state of the missing action.

Crucially, the model does not attempt to generate the missing video frames pixel by pixel. Instead, it predicts the abstract evolution of the scene over time. If the video depicts a person throwing a ball, the predictor network forecasts the continuous embedding that represents the trajectory and physics of the throw. By training exclusively on these latent temporal transitions, the model learns the intuitive physics of the environment, including gravity, momentum, and object permanence, which are impossible to extract from static images or autoregressive text.

8.3. Audio and Multimodal Expansions A comprehensive world model cannot rely on a single sensory input; physical reality is inherently multimodal. The architectural principles of latent space prediction have naturally expanded to encompass auditory data and synchronized multimodal streams. The Audio Joint Embedding Predictive Architecture applies a similar masking strategy to audio spectrograms, forcing the system to predict the abstract representations of missing auditory temporal segments rather than raw acoustic waveforms. More significantly, recent advancements have focused on combining these modalities into unified architectures, where the system must learn joint representations across both audio and visual inputs. In a multimodal setting, a model might utilize a visible video context to predict the latent representation of an accompanying masked audio track, or vice versa. This cross-modal prediction forces the architecture to understand the causal relationship between different sensory phenomena, such as associating the visual impact of a dropped object with the specific acoustic signature of its crash. By grounding its predictive objectives in multiple, synchronized data streams, the architecture builds a far more robust and cohesive internal simulation of physical reality.

9. Applications in Robotics and Autonomous Systems

9.1. Navigating Continuous Physical Environments The theoretical advantages of latent space prediction translate directly into profound capabilities for physical robotics. Traditional robotic control systems heavily rely on either rigidly programmed kinematics, which fail in dynamic environments, or deep reinforcement learning policies trained on massive amounts of trial-and-error data, which struggle to generalize to novel situations. When a robot powered by a standard autoregressive or purely reactive model encounters an unfamiliar obstacle, it lacks the intuitive physics required to adapt safely. Conversely, a robot equipped with a Joint Embedding Predictive Architecture functions using an internal world model. It processes its continuous physical environment through its sensors, compresses that sensory data into a latent state, and uses its predictor network to simulate the consequences of various potential movements. Because the architecture understands abstract physical concepts like mass, momentum, and spatial geometry, the robot can mentally navigate the environment and evaluate outcomes before executing a physical action. This allows for highly adaptive navigation in unstructured spaces, such as search and rescue operations or domestic environments, where the physical terrain is unpredictable and constantly shifting.

9.2. Industrial Automation and Real-Time Decision Making In the domain of industrial automation, efficiency, precision, and safety are paramount. Manufacturing environments involve complex, multi-step processes where robotic manipulators must interact with varying materials, handle delicate components, and respond instantly to anomalies on an assembly line. Generative models are often unsuited for this task due to their high latency and reliance on pixel-level processing. The predictive discriminative approach of latent world models, which AMI Labs is specifically targeting for factory floor "cobots" and process control, offers a critical solution. By operating entirely within a compressed mathematical space, the architecture requires significantly less computational overhead during inference, enabling real-time decision making. If a robotic arm is tasked with sorting irregular objects, its world model can instantly simulate the physics of grasping a uniquely shaped item, predicting slippage or instability before the gripper even makes contact. Furthermore, because the model naturally filters out visual noise, such as changing factory lighting or dust on a camera lens, it maintains high reliability. This capability transforms industrial robots from rigid, repetitive machines into highly adaptable agents capable of continuous real-time problem solving in "messy" high-dimensional environments.

9.3. Simulation Engines for Autonomous Driving Perhaps the most critical application for artificial intelligence world models is autonomous driving, where the "HanoiWorld" and "AD-L-JEPA" frameworks have demonstrated the power of latent prediction for LiDAR and video streams. Navigating a vehicle through dense urban traffic is a quintessential example of a highly

dynamic, unpredictable physical environment. Current systems often struggle with edge cases—rare events that fall outside their training distribution. Attempting to solve this by scaling autoregressive video generation is a computational dead end, as the system would need to render every pixel trajectory of every pedestrian and vehicle in real time. A latent world model, however, acts as a highly efficient internal simulation engine. The autonomous vehicle encodes the current traffic state into a latent vector and simulates multiple future scenarios based on different driving actions, such as braking or swerving. The model predicts the abstract semantic flow of traffic, entirely ignoring irrelevant details like the exact reflections on a neighboring car's windshield. By evaluating these simulated futures against a safety cost function, the vehicle can execute complex, long-horizon planning and navigate edge cases with a level of causal reasoning previously exclusive to human drivers.

10. Integrating System 2 Thinking in Artificial Intelligence

10.1. Defining Fast versus Slow Thinking in Neural Networks The conceptual framework of fast and slow thinking, originally popularized by Daniel Kahneman, provides a profound analogy for the current evolution of machine learning architectures. Standard autoregressive models operate almost entirely within the paradigm of System 1, or fast thinking. They process an input context and immediately react, generating the next most statistically likely token or pixel in a continuous, uninterrupted flow. In these systems, there is no pause for deliberation, no internal drafting process, and no structural mechanism to evaluate the long-term logical consistency of an output before it is produced. Conversely, System 2 thinking is deliberate, analytical, and requires the mental simulation of various possibilities before a final decision is reached. By separating the predictive mechanism from the generation of raw sensory outputs, the Joint Embedding Predictive Architecture finally provides a foundational structure for slow thinking in neural networks. It allows an artificial agent to halt its immediate reaction, transition into a compressed mathematical space, and dedicate computational time to internal planning rather than mere reactive pattern matching.

10.2. Logical Reasoning and Counterfactual Simulation The capacity for slow deliberation directly enables complex logical reasoning through the use of counterfactual simulation. When a traditional generative model is presented with a novel logical puzzle or a complex physical scenario, it relies entirely on whether a similar sequential structure existed within its training corpus. If the specific scenario is entirely novel, the model frequently fails because it cannot test the validity of its own logic. A latent world model approaches this challenge through an entirely different mechanism. When faced with a decision or a complex environment, the architecture can utilize its predictor network to simulate multiple counterfactual realities. It mathematically poses the question of what would happen if a specific action is taken versus an alternative

action. By projecting these different choices forward in time through the continuous embedding space, the system can compare the divergent outcomes against a predefined objective or cost function. This ability to explore parallel, simulated futures allows the artificial intelligence to deduce the optimal course of action through structured causal reasoning rather than blind statistical memorization.

10.3. Escaping the Reactive Feedback Loop One of the most persistent and structural flaws in autoregressive generation is the system's inherent vulnerability to the reactive feedback loop. Because these models generate outputs sequentially and immediately feed those outputs back into their own context window for the next step, a single hallucination or logical error early in a sequence rapidly cascades into total catastrophic failure. The model is forced to justify and build upon its own mistakes because it fundamentally lacks a global view of the overarching trajectory. Artificial intelligence world models escape this reactive trap by fundamentally decoupling the planning phase from the execution phase. In a system powered by the Joint Embedding Predictive Architecture, an agent simulates an entire trajectory of actions in the abstract latent space before ever committing to a single action in physical reality.

If a simulated path leads to a high-energy, undesirable state or a logical dead end, the agent simply discards that mental trajectory and searches for a superior route. This paradigm shift from greedy, step-by-step generation to global, objective-driven planning ensures that the artificial intelligence remains robust, capable of self-correction, and strictly aligned with its ultimate goals.

11. Challenges and Scalability

11.1. Hardware and Training Complexities While the Joint Embedding Predictive Architecture offers significant computational advantages during inference by eliminating the need to decode raw pixels, the training phase remains a formidable engineering challenge. Developing a robust world model requires the system to ingest and compress vast amounts of multimodal data to accurately map the underlying physics and semantic rules of reality. As AMI Labs scales this architecture from static images to continuous, high-definition video and synchronized sensory streams, the memory requirements expand. Processing massive spatiotemporal tubes through deep encoder-only transformers demands immense memory bandwidth and highly optimized distributed training across large-scale clusters. For perspective, training a high-capacity variant like ViT-Huge on a large-scale dataset typically requires dozens of high-end GPUs, such as the NVIDIA A100 or H100. Furthermore, balancing the learning dynamics between the context encoder, the target encoder, and the predictor network requires precise hyperparameter tuning. If the predictor network learns too quickly, it may overpower the encoder's ability to extract

meaningful features; if the encoder updates too slowly, the system fails to capture complex temporal dynamics.

11.2. Representation Collapse and Mitigation Strategies The most persistent mathematical challenge in training joint embedding architectures is the phenomenon of representation collapse. Because the predictive objective function seeks to minimize the distance between the predicted state and the actual encoded target state, typically expressed as L = ||s_y - s_hat_y||^2, the neural networks can discover a trivial mathematical loophole. Instead of learning complex semantic features, the encoder network can simply map every possible input to the exact same constant latent vector, such that s_y = c. Consequently, the predictor network easily outputs that same constant c, driving the loss to zero without the system actually learning anything. To mitigate this, JEPA employs an asymmetric architecture and a stop-gradient mechanism. The target encoder's weights are updated using an Exponential Moving Average (EMA) of the context encoder's weights, preventing the target representations from becoming a static target. Newer iterations, such as C-JEPA, also integrate Variance-Invariance-Covariance Regularization (VICReg) to explicitly ensure the variance of the embeddings remains high, mathematically forbidding the model from collapsing into a single constant point.

11.3. The Future Trajectory of Joint Embedding Architectures The future trajectory of latent world models points toward hierarchical planning and deeper integration with active agency. Current architectures focus primarily on short-term temporal scales, forecasting events a few seconds into the future. However, true intelligence requires hierarchical temporal abstraction. Future iterations of the Joint Embedding Predictive Architecture are expected to feature nested predictor networks where lower levels predict immediate physical interactions, while higher levels predict long-term goal completions over minutes or hours. Furthermore, moving these models from passive observation to active "embodied" agency requires coupling the latent variable z with complex robotic actuation policies. By allowing an autonomous agent to continuously query its internal world model with proposed physical actions, evaluate the outcomes in latent space, and execute the optimal path in real time, the field moves closer to artificial general intelligence grounded in the reality of the physical world rather than mere statistical text generation. AMI Labs' massive funding suggests that the next phase will involve collecting proprioceptive and multimodal data at an unprecedented scale to bridge the gap between virtual simulations and real-world physical complexity.

12. Conclusion

12.1. Summary of Architectural Contrasts The transition from autoregressive generative models to predictive discriminative architectures marks a profound philosophical and structural pivot in the pursuit of machine intelligence. Traditional large language models and pixel-level generative systems have demonstrated extraordinary capabilities in syntactic fluency and pattern replication. However, their fundamental reliance on predicting discrete data units within a high-dimensional input space binds them to computational inefficiency and a distinct lack of physical grounding. By obsessing over the exact reconstruction of stochastic noise and irrelevant micro-details, these systems fail to capture the causal mechanics of the physical world. In stark contrast, the Joint Embedding Predictive Architecture bypasses these limitations by operating entirely within a continuous, abstract latent space. By compressing sensory data into semantic representations and predicting the structural evolution of those states, this alternative architecture filters out unpredictable noise, embraces aleatoric uncertainty, and successfully models the underlying physical dynamics of its environment.

12.2. Final Thoughts on the Future of World Models The emergence of artificial intelligence world models, accelerated by the massive 1.03 billion dollar seed funding of Advanced Machine Intelligence Labs in early 2026, signals a definitive departure from purely reactive System 1 systems toward deliberative System 2 agents. The true value of machine intelligence in the physical world does not lie in hallucinating plausible visual media, but in safely and autonomously navigating complex, continuous environments. By equipping artificial agents with an internal cognitive simulator, world models enable systems to project counterfactual realities, evaluate long-term consequences, and execute objective-driven planning before committing to a physical action. This shift from statistical memorization to cognitive simulation is the essential catalyst for deploying robust autonomous systems in robotics, industrial automation, and healthcare, where an intuitive grasp of physics and causality is non-negotiable.

12.3. Call for Future Research While the theoretical foundation of the Joint Embedding Predictive Architecture is compelling, the realization of fully capable world models requires extensive future research and structural refinement. The academic and industrial communities must prioritize the development of hierarchical temporal abstractions, enabling models to plan across extended time horizons rather than merely forecasting immediate physical transitions. Furthermore, significant research is needed to seamlessly integrate active reinforcement learning policies with latent predictive states, closing the loop between internal cognitive simulation and external physical actuation. As Yann LeCun and AMI Labs move toward commercializing these models for robotics and wearable devices, addressing the hardware bottlenecks associated with training massive multimodal embedding architectures remains a critical frontier. Reallocating research efforts toward objective-driven, latent space architectures

offers the most promising trajectory toward achieving true artificial general intelligence grounded in the reality of the physical world.

13. References

Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., and Ballas, N. (2023). Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 21019.

Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., Assran, M., and Ballas, N. (2024). V-JEPA: Latent Video Prediction for Visual Representation Learning. arXiv preprint arXiv:2404.08471.

Grill, J. B., Strub, F., Altche, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., and Valko, M. (2020). Bootstrap your own latent: A new approach to self-supervised learning. Advances in Neural Information Processing Systems, 33, 21271-21284.

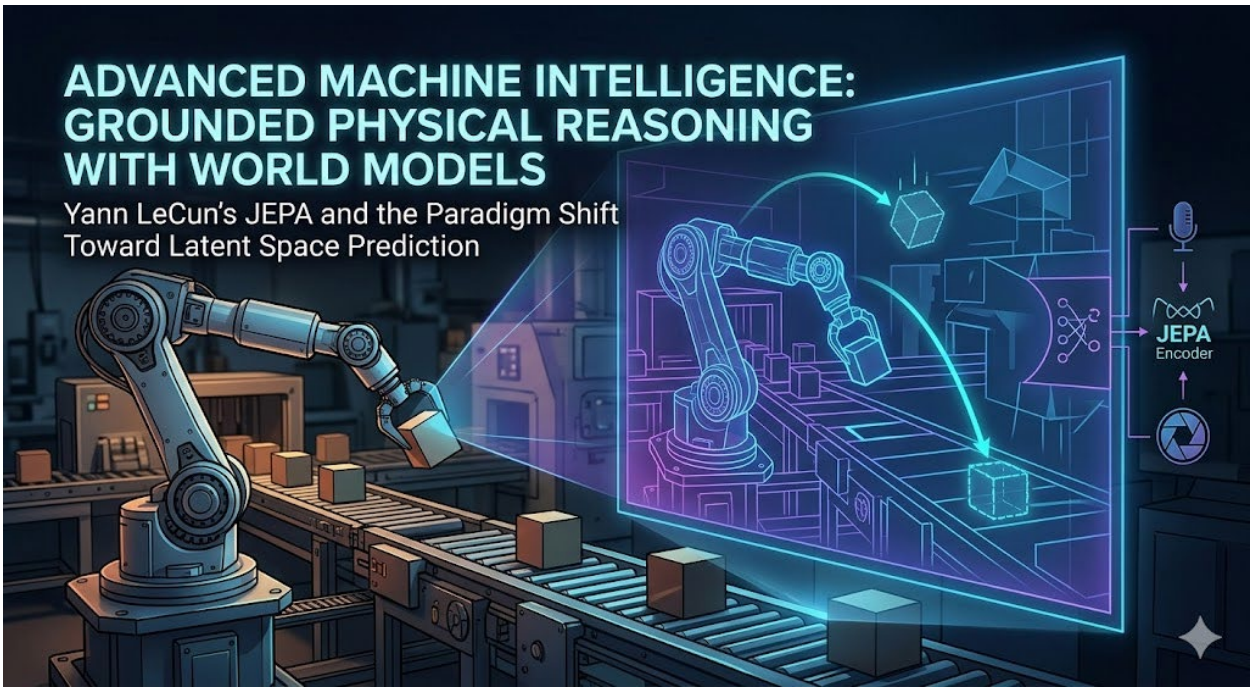
LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., and Huang, F. (2006). A Tutorial on Energy-Based Learning. Predicting Structured Data, 1(0), 191-246.

LeCun, Y. (2022). A Path Towards Autonomous Machine Intelligence Version 0.9.2. OpenReview.

Reuters (2026). Ex-Meta AI chief Yann LeCun's AMI raises 1.03 billion for alternative AI approach. Published March 10, 2026.

Shorooq (2026). Shorooq invests in AMI Labs as AI startup hits 3.5 billion valuation. Wamda, News Release March 11, 2026.

The author has published over 5,000 AI-assisted papers at: https://www.researchgate.net/profile/Douglas-Youvan/research . Google Scholar has indexed these preprints, and if you want a particular reference, the AI mode of a Google search (Gemini) has efficiently catalogued these preprints.


ADVANCED MACHINE INTELLIGENCE: GROUNDED PHYSICAL REASONING WITH WORLD MODELS
Yann LeCun's JEPA and the Paradigm Shift Toward Latent Space Prediction
JEPA
Encoder